ROBUST MULTIMODAL SENTIMENT ANALYSIS VIA CALIBRATED DYNAMIC FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Sentiment Analysis (MSA), which integrates linguistic, acoustic, and visual modalities, is crucial for understanding human emotions. Since modality quality often varies in real-world scenarios, dynamic fusion methods have been introduced to adaptively weight modalities according to their estimated reliability. However, existing methods still suffer from two key limitations: they often confuse degradation caused by intrinsic noise with that caused by insufficient optimization, inducing erroneous modality modulation. Besides, they usually rely on simple probabilistic assumptions that cannot capture the complex, multi-peak structure of real modality features. To address these issues, we propose a Calibrated Dynamic Fusion (CDF) framework for robust MSA. Specifically, we introduce Degradation Source Rectification (DSR) to reduce optimization-induced uncertainty bias, preventing informative but under-optimized modalities from being unduly downweighted. We further propose Multi-Peak Uncertainty Modeling (MPUM) to capture heterogeneous local patterns in modality representations and provide more faithful uncertainty estimates under complex feature distributions. The calibrated uncertainty estimates are then used to assess modality reliability and guide sample-adaptive dynamic fusion.

1 INTRODUCTION

Multimodal sentiment analysis (MSA) combines complementary language, acoustic, and visual cues for sentiment understanding Yang et al. (2023); A.V. et al. (2024). However, modality reliability varies across samples: speech may convey clear emotional cues while visual signals are weakened by occlusion or neutral expressions. Thus, direct concatenation or uniform weighting may introduce unreliable information and impair fusion.

To this end, numerous dynamic MSA methods have been proposed to adaptively adjust modality contributions according to sample-dependent modality reliability. These methods can be broadly categorized into three types. The first type focuses on attention-based dynamic fusion, which learns input-dependent modality weights from cross-modal interactions Feng et al. (2024). However, such methods mainly rely on deterministic feature reweighting, limiting their ability to capture complex cross-modal dependencies and subtle variations in modality quality. To improve the flexibility of dynamic fusion, the second type introduces structure-enhanced mechanisms, including graph reasoning, dynamic experts, feature disentanglement, and sample-specific modality selection Liu et al. (2025); Yang et al. (2026); Ali & Hughes (2023); Fang et al. (2025); Wang et al. (2025). Although these methods capture richer cross-modal relationships, they still operate primarily on deterministic point representations, making it difficult to explicitly quantify modality reliability under noisy or ambiguous conditions. This motivates the third type of research based on probabilistic representation learning. It models each modality as a probability distribution and derives uncertainty from distributional statistics to guide dynamic fusion Huang et al. (2025); Tellamekala et al. (2023); Zhu et al. (2025); Hu et al. (2025).

Despite recent progress, existing methods still suffer from two major limitations. First, they implicitly attribute representation degradation to noise corruption and accordingly assign lower fusion weights to degraded modality representations. However, in practice, representation degradation can also stem from non-noise factors. For example, cross-modal optimization competition may leave harder-to-learn modalities insufficiently optimized Peng et al. (2022); Huang et al. (2022); Poria

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

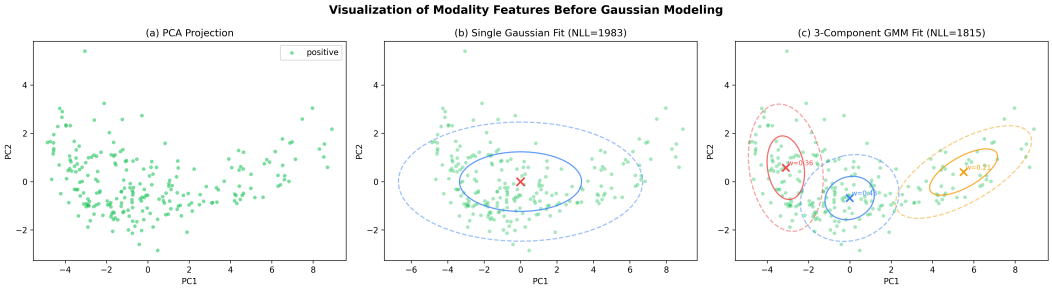


Figure 1: PCA visualization of textual features for positive samples from CMU-MOSI. The three-component GMM achieves a lower negative log-likelihood than the single Gaussian (1815 vs. 1983), suggesting that a mixture distribution provides a more flexible fit to the heterogeneous feature patterns.

et al. (2020), so that they appear unreliable even when they are not severely corrupted. Down-weighting such modalities may further suppress informative cues. Second, existing probabilistic methods typically model each modality with a simple single Gaussian, and summarize uncertainty using one overall variance. This overlooks the heterogeneous local patterns of each modality. As shown in Fig. 1, the textual features exhibit multiple local structures. Although a single Gaussian can provide a coarse fit to the overall distribution, it yields a higher negative log-likelihood than the three-component GMM (1983 vs. 1815). This suggests that the single-Gaussian assumption cannot adequately capture heterogeneous modality patterns.

To address these challenges, we propose a Calibrated Dynamic Fusion (CDF) framework for robust multimodal sentiment analysis. CDF consists of two complementary modules: Degradation Source Rectification (DSR) and Multi-Peak Uncertainty Modeling (MPUM). DSR introduces unimodal regularization to strengthen unimodal learning and alleviate under-optimization caused by modality competition. This reduces the uncertainty bias arising from insufficient learning. Since this additional supervision can conflict with the multimodal objective, we further employ a gradient decoupling strategy, where multimodal loss updates the fusion and prediction modules, while unimodal losses optimize the modality encoders. Based on the rectified representations, MPUM models each modality with a sample-conditional Gaussian mixture and models the modality through the within-component aleatoric uncertainty and predictive uncertainty. This captures expert disagreement and component assignment ambiguity. Finally, extensive experiments on multiple multimodal benchmarks demonstrate the effectiveness of CDF.

Overall, our contributions are summarized as follows:

- We propose a Calibrated Dynamic Fusion (CDF) framework for robust multimodal sentiment analysis, which improves dynamic fusion through calibrated modality reliability estimation.
- We propose a Degradation Source Rectification module that rectifies the influence of optimization-induced degradation on noise-induced degradation estimation. This reduces the erroneous suppression of hard-to-learn modalities and yields cleaner modality representations for uncertainty estimation.
- We propose Multi-Peak Uncertainty Modeling to capture heterogeneous local patterns and characterize feature uncertainty through complementary aleatoric and predictive components. This enables more accurate uncertainty estimation for complex modality representations.

2 Related Work

2.1 Multimodal Representation and Fusion

MSA methods exploit complementary information through cross-modal interactions. TFN and LMF capture high-order interactions via tensor-based fusion Zadeh et al. (2017); Liu et al. (2018), while

MISA separates modality-invariant and modality-specific representations Hazarika et al. (2020). MulT learns pairwise cross-modal dependencies Tsai et al. (2019). More recent methods account for modality asymmetry: DLF emphasizes the language modality Wang et al. (2025), and MODS selects a primary modality for each sample Yang et al. (2026). These methods improve cross-modal interaction but determine modality contributions without explicitly modeling uncertainty.

2.2 Uncertainty-Aware Multimodal Fusion

Uncertainty-aware methods assess modality reliability through probabilistic modeling. COLD Fusion uses latent variance to evaluate temporal informativeness Tellamekala et al. (2023), and UEFN estimates decision-level uncertainty using Dirichlet evidence Sun et al. (2024). LDDU and P-RMF derive uncertainty from probabilistic latent representations Huang et al. (2025); Zhu et al. (2025). However, restrictive distributional assumptions can limit the representation of heterogeneous local patterns, compromising reliability estimation. MPUM addresses this limitation by modeling multi-peak structure within modality representations.

2.3 Modality Imbalance and Optimization

Joint training can favor faster-learning or more predictive modalities, leaving others insufficiently optimized. Existing methods mitigate this imbalance through learning-rate modulation, gradient balancing, and modality-specific optimization Peng et al. (2022); Huang et al. (2022); Poria et al. (2020); Li et al. (2023a); Wei et al. (2025b; 2024). For uncertainty-aware fusion, sufficient unimodal learning is also essential: under-optimized representations can bias reliability estimates and cause informative modalities to be downweighted. DSR addresses this issue through unimodal regularization, complementing MPUM by improving the representations from which uncertainty is estimated.

3 Methodology

3.1 Problem Formulation and Framework Overview

We consider an MSA dataset $\mathcal{D} = \{(\{X_i^m\}_{m\in\mathcal{M}}, y_i)\}_{i=1}^N$, where N denotes the number of samples, $\mathcal{M} = \{l, a, v\}$ denotes the language, acoustic, and visual modalities, and y_i is the sentiment label of the i-th sample. For each modality m, the input sequence X_i^m is encoded by a modality-specific encoder:

$$H_i^m = E_m(X_i^m), \qquad h_i^m = \mathrm{Pool}(H_i^m), \tag{1}$$

where H_i^m and h_i^m denote the sequence-level and utterance-level representations, respectively.

Dynamic fusion assigns a sample-specific weight ω_i^m to each modality according to its estimated reliability, yielding $\{\omega_i^m\}_{m\in\mathcal{M}}$. The weighted modality representations are aggregated into a proxy representation and further refined through cross-modal interaction:

$$p_i = G_p\left(\omega_i^l h_i^l \oplus \omega_i^a h_i^a \oplus \omega_i^v h_i^v\right), \qquad \hat{y}_i = P_f\left(\mathrm{Attn}\left(p_i, h_i^l, h_i^a, h_i^v\right)\right), \qquad \mathcal{L}_{\text{fuse}} = \ell(\hat{y}_i, y_i) \tag{2}$$

where $\oplus$ denotes concatenation, $G_p(\cdot)$ is the proxy projection network, $P_f(\cdot)$ is the prediction head and ℓ is the cross-entropy loss.

The effectiveness of dynamic fusion therefore depends on whether $\{\omega_i^m\}_{m\in\mathcal{M}}$ accurately reflects the reliability of each modality for the i-th sample. To this end, as illustrated in Fig. 2, CDF introduces the Degradation Source Rectification (DSR) and Multi-Peak Uncertainty Modeling (MPUM). DSR reduces optimization-induced bias in modality representations, while MPUM estimates modality uncertainty under heterogeneous feature distributions.

3.2 Degradation Source Rectification

As discussed, dynamic fusion requires uncertainty estimates to accurately reflect input-dependent modality degradation. However, conventional joint training may under-optimize difficult modalities, causing their weak representations to receive excessive uncertainty. To address this issue, we propose DSR, which disentangles noise-induced degradation from optimization-induced degradation.

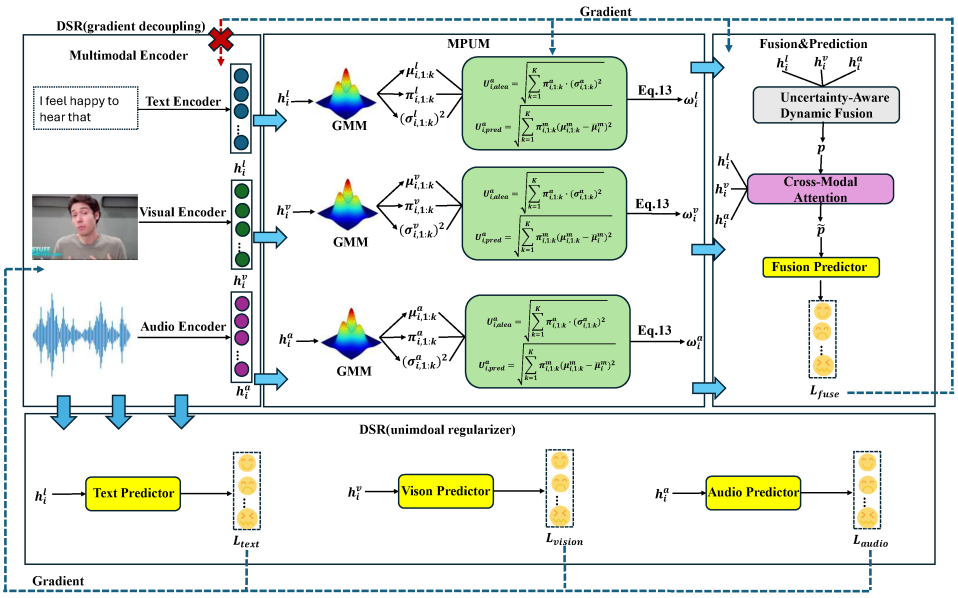


Figure 2: Overall architecture of the proposed Calibrated Dynamic Fusion (CDF) framework. DSR reduce the optimization basis through supervision and gradient decoupling, while MPUM models heterogeneous modality distributions and estimates calibrated uncertainty for dynamic fusion.

First, for each modality, we introduce an independent prediction head P_m:

$$\hat{y}_i^m = P_m(h_i^m), \qquad \mathcal{L}_{\text{uni}} = \frac{1}{\mathcal{M}} \sum_{m \in \mathcal{M}} \ell(\hat{y}_i^m, y_i), \tag{3}$$

where $\hat{y}_i^m$ denotes the prediction of the i-th sample using only modality m, and $\ell(\cdot, \cdot)$ denotes the sentiment prediction loss. By directly supervising each modality with the task label, $\mathcal{L}_{\text{uni}}$ improves modality-specific representations and reduces the uncertainty bias caused by insufficient learning.

However, allowing the unimodal and multimodal objectives to jointly update the modality encoders may reintroduce optimization competition. We therefore employ gradient decoupling by applying stop-gradient before uncertainty modeling and multimodal fusion:

$$\widetilde{h}_i^m = \text{sg}(h_i^m), \qquad i = 1, \ldots, N, \quad m \in \mathcal{M}, \tag{4}$$

where $\widetilde{h}_i^m$ preserves the forward value of h_i^m, while $\text{sg}(\cdot)$ blocks the backward gradient from downstream multimodal objectives.Under this decoupled optimization scheme, $\mathcal{L}_{\text{uni}}$ updates the modality encoders and unimodal prediction heads, whereas $\mathcal{L}_{\text{fuse}}$ updates the multimodal fusion branch. Consequently, the modality representations used for uncertainty estimation are learned without interference from multimodal optimization.

Notably, this does not weaken cross-modal interaction. The detached representations preserve their forward values and are still jointly processed by the fusion modules. Meanwhile, these modules remain optimized by the multimodal objectives. Therefore, cross-modal dependencies can still be effectively learned within the fusion stage.

3.3 Multi-Peak Uncertainty Modeling

While DSR mitigates optimization-induced bias, reliable dynamic fusion still requires accurate uncertainty modeling. Existing methods approximate each sample with a single Gaussian, which collapses multiple plausible sentiment patterns into a coarse distribution. To address this limitation, we propose Multi-Peak Uncertainty Modeling (MPUM), which models the sentiment distribution with

a Gaussian mixture:

$$q_i^m\left(y_i \mid \widetilde{h}_i^m\right)=\sum_{k=1}^{K}\pi_{i,k}^m\mathcal{N}\left(y_i;\mu_{i,k}^m,\left(\sigma_{i,k}^m\right)^2\right), \tag{5}$$

where $\pi_{i,k}^m \geq 0$ and $\sum_{k=1}^{K}\pi_{i,k}^m = 1$. Each of the K components represents a local sentiment prediction pattern: $\mu_{i,k}^m$ denotes its prediction center, $\left(\sigma_{i,k}^m\right)^2$ describes its within-pattern dispersion, and $\pi_{i,k}^m$ indicates its plausibility for the current sample.

To determine the parameters in Eq. 5, MPUM first projects $\widetilde{h}_i^m$ into a compact latent representation and then generates the mixture parameters through a modality-specific router and K Gaussian experts:

$$\begin{aligned} z_i^m &= F_m(\widetilde{h}_i^m), \qquad \boldsymbol{\pi}_i^m = \mathrm{Entmax}_{1.5}\left(R_m(z_i^m)/\tau\right), \\ \mu_{i,k}^m &= G_{m,k}^{\mu}(z_i^m), \qquad \left(\sigma_{i,k}^m\right)^2 = \sigma_{\min}^2 + (\sigma_{\max}^2-\sigma_{\min}^2)\cdot\mathrm{Sigmoid}\left(G_{m,k}^{\sigma}(z_i^m)\right) \end{aligned} \tag{6}$$

Here, z_i^m denotes the latent representation of modality m for sample i, obtained through the modality-specific projection network F_m. R_m is the routing network, while $G_{m,k}^{\mu}$ and $G_{m,k}^{\sigma}$ denote the mean and variance heads of the k-th Gaussian component, respectively. In practice, these mappings are implemented as lightweight two-layer MLPs, introducing negligible computational overhead. Entmax converts the routing logits into sparse and normalized mixture probabilities, with τ controlling the routing sharpness. For variance estimation, the unconstrained output of $G_{m,k}^{\sigma}$ is first mapped by Sigmoid to $(0,1)$ and then linearly rescaled to $(\sigma_{\min}^2,\sigma_{\max}^2)$. This bounded parameterization guarantees valid positive variances and prevents both variance collapse and excessively diffuse Gaussian components.

For a mini-batch of B samples, the conditional mixture is learned using the negative log-likelihood:

$$\mathcal{L}_{\mathrm{nll}} = -\frac{1}{B|\mathcal{M}|}\sum_{i=1}^{B}\sum_{m\in\mathcal{M}}\log q_i^m\left(y_i \mid \widetilde{h}_i^m\right). \tag{7}$$

Although $\mathcal{L}_{\mathrm{nll}}$ fits the overall mixture distribution, it lacks explicit supervision for component selection. Consequently, the router may assign high probability to a component that poorly explains the ground-truth sentiment. To guide component selection, we quantify how strongly each Gaussian component supports the ground-truth label and use these scores to supervise the relative routing weights:

$$\xi_{i,k}^m = \frac{\mathcal{N}\left(y_i;\mu_{i,k}^m,\left(\sigma_{i,k}^m\right)^2\right)}{\sum_{j=1}^{K}\mathcal{N}\left(y_i;\mu_{i,j}^m,\left(\sigma_{i,j}^m\right)^2\right)}. \tag{8}$$

Here, $\xi_{i,k}^m$ measures how well the k-th component explains the ground-truth sentiment of the i-th sample.

We then align the routing probabilities with the label-guided target:

$$\mathcal{L}_{\mathrm{route}} = \frac{1}{B|\mathcal{M}|}\sum_{i=1}^{B}\sum_{m\in\mathcal{M}}\mathrm{KL}\left(\mathrm{sg}(\boldsymbol{\xi}_i^m)\,\|\,\boldsymbol{\pi}_i^m\right), \tag{9}$$

where $\mathrm{sg}(\cdot)$ prevents the routing loss from updating the Gaussian experts. In this way, $\mathcal{L}_{\mathrm{route}}$ explicitly guides the router toward components that better explain the ground-truth sentiment, rather than relying solely on the indirect supervision of the mixture likelihood.

The complete MPUM objective is defined as

$$\mathcal{L}_{\mathrm{MPUM}} = \lambda_{\mathrm{nll}}\mathcal{L}_{\mathrm{nll}} + \lambda_{\mathrm{route}}\mathcal{L}_{\mathrm{route}} \tag{10}$$

Here, $\mathcal{L}_{\mathrm{nll}}$ fits the overall conditional distribution, $\mathcal{L}_{\mathrm{route}}$ supervises sample-wise component selection.

3.4 Uncertainty-Guided Dynamic Fusion

Based on the mixture distribution learned by MPUM, we estimate modality uncertainty from within-pattern dispersion and cross-pattern ambiguity.

First, to calculate the uncertainty within each sentiment pattern. We extend the variance-based uncertainty of a single Gaussian to the mixture distribution by weighting the variance of each component:

$$U^m_{i,\text{alea}} = \sum_{k=1}^{K} \pi^m_{i,k} \left(\sigma^m_{i,k}\right)^2. \tag{11}$$

This aleatoric uncertainty measures the expected dispersion within plausible sentiment patterns.

However,small within-pattern variances do not necessarily imply reliable modality predictions. Different components may each produce concentrated predictions while disagreeing on sentiment polarity or intensity. For example, two components may confidently predict positive and negative sentiment, respectively. Their individual variances remain small, but their conflicting predictions reveal ambiguity that within-pattern dispersion alone cannot capture.

To account for this ambiguity, we introduce predictive uncertainty to quantify disagreement among component predictions. Let $\bar{\mu}^m_i = \sum_{k=1}^{K} \pi^m_{i,k}\mu^m_{i,k}$ denote the mixture prediction. We define

$$U^m_{i,\text{pred}} = \sum_{k=1}^{K} \pi^m_{i,k} \left(\mu^m_{i,k} - \bar{\mu}^m_i\right)^2. \tag{12}$$

Here, each squared deviation measures how far a component prediction departs from the mixture prediction, while the routing weight $\pi^m_{i,k}$ scales its contribution according to the component's relevance to the current sample. This quantity is zero when all active components agree and becomes large when components with substantial routing weights produce divergent predictions.

Finally, the overall modality uncertainty is defined as

$$U^m_i = \lambda_a U^m_{i,\text{alea}} + \lambda_p U^m_{i,\text{pred}}, \tag{13}$$

where λ_a and λ_p balance within-pattern dispersion and cross-pattern predictive ambiguity, respectively.

Thus, we calculate the sample-specific fusion weights from the estimated modality uncertainty as follows:

$$r^m_i = \left(1 + U^m_i\right)^{-\beta}, \qquad \omega^m_i = \frac{r^m_i}{\sum_{j\in\mathcal{M}} r^j_i}, \tag{14}$$

where β controls the sensitivity to uncertainty. The resulting weights are used in Eq. equation 2 for sample-adaptive fusion.

3.5 Total Loss

The overall objective of CDF is

$$\mathcal{L} = \alpha\mathcal{L}_{\text{uni}} + \mathcal{L}_{\text{fuse}} + \mathcal{L}_{\text{MPUM}}, \tag{15}$$

where α controls the strength of unimodal supervision. Under gradient decoupling, $\mathcal{L}_{\text{uni}}$ updates the modality encoders and unimodal prediction heads, while $\mathcal{L}_{\text{fuse}}$ optimizes the downstream fusion branch. $\mathcal{L}_{\text{MPUM}}$ directly trains the probabilistic mixture model.

4 Experiments

4.1 Experimental Setup

We evaluate CDF on CMU-MOSI Zadeh et al. (2016) and CMU-MOSEI Bagher Zadeh et al. (2018) and conduct robustness experiments on MVSA-Single Niu et al. (2016). Detailed dataset statistics, feature preprocessing, and complete hyperparameter settings are provided in the appendix.

Table 1: Performance comparison on the CMU-MOSI and CMU-MOSEI datasets. Bold and underlined values denote the best and second-best results, respectively. "–" indicates an unreported result.

Methods	CMU-MOSI						CMU-MOSEI					
	Acc-7↑	Acc-5↑	Acc-2↑	F1↑	Corr↑	MAE↓	Acc-7↑	Acc-5↑	Acc-2↑	F1↑	Corr↑	MAE↓
Graph-MFN	34.64	38.63	78.35	78.35	0.6490	0.9560	51.37	52.69	83.48	83.43	0.7130	0.5750
MulT	40.00	42.68	83.00	82.00	0.6980	0.8710	51.80	54.18	82.50	82.30	0.7030	0.5800
PMR	40.60	–	83.60	83.60	–	–	52.50	–	83.60	83.40	–	–
MISA	41.37	47.08	83.54	83.58	0.7780	0.7770	52.05	53.63	84.67	84.66	0.7520	0.5580
MAG-BERT	43.62	–	84.43	84.61	0.7810	0.7270	52.67	–	84.82	84.71	0.7550	0.5430
DMD	46.06	–	83.23	83.29	–	0.7520	52.78	–	84.62	84.62	–	0.5430
DLF	44.02	51.17	83.38	83.45	0.7893	0.7571	52.56	54.26	83.96	84.04	0.7583	0.5528
LDDU	47.52	52.79	84.32	84.25	0.7856	0.7551	53.72	54.02	84.55	84.29	0.7482	0.5548
MODS	49.27	–	85.83	85.96	0.7980	0.6880	54.32	–	85.88	86.14	0.7720	0.5270
Ours (CDF)	**49.98**	**54.81**	**87.04**	**87.02**	**0.8042**	**0.6835**	**55.66**	**56.04**	**87.27**	**87.34**	**0.7892**	**0.5216**

Table 2: Comparative results under the Robust Multimodal Learning settings. ϵ denotes the noise strength.

Method	MVSA-Single				
	Clean	Gaussian@ ϵ		Salt-Pepper@ ϵ	
		5.0	10.0	5.0	10.0
EAU	79.15	73.34	61.78	73.69	60.46
ECML	76.83	71.28	61.03	72.13	61.04
NLC	73.79	65.39	58.98	66.64	57.28
IPRM	75.84	71.25	60.69	70.13	58.26
ARL	75.76	70.89	60.74	70.49	59.82
CPSC	80.07	74.12	63.32	73.95	61.27
CDF (Ours)	**80.92**	**75.07**	**64.66**	**75.25**	**63.09**

Table 3: Expected calibration error (ECE) on MVSA-Single.

Method	Text ECE↓	Image ECE↓
Single Gaussian	25.98	27.45
Single Gaussian + DSR	26.02	27.55
MPUM	20.46	17.69
MPUM + DSR (CDF)	**19.92**	**12.64**

4.1.1 Main results.

Table 1 shows that CDF ranks first on all six metrics across CMU-MOSI and CMU-MOSEI. Compared with MODS, CDF improves Acc-7, Acc-2, and F1 by $0.71/1.21/1.06$ points on CMU-MOSI and $1.34/1.39/1.20$ points on CMU-MOSEI, respectively. These gains cover both sentiment polarity and intensity. Higher Corr and lower MAE also indicate improved regression accuracy. The results suggest that representation rectification and calibrated uncertainty help preserve discriminative information and guide reliability-aware fusion.

4.1.2 Robustness results.

Real-world MSA often involves modality-specific corruption, under which unreliable information may adversely affect fusion if modality quality changes are not properly reflected in the weighting process. We therefore inject Gaussian and salt-and-pepper noise at increasing strengths to examine whether CDF can adapt to degraded inputs. As shown in Table 2, CDF performs best under all clean and corrupted settings, with its margin over CPSC increasing from 0.85 points on clean data to 1.44 and 1.82 points under the strongest Gaussian and salt-and-pepper noise, respectively. These results show that calibrated uncertainty reduces the contribution of degraded evidence through sample-dependent weighting.

4.2 Ablation Studies

4.2.1 Overall component ablation.

Table 4 verifies the two core designs of CDF. MPUM raises Acc_7 from 33.83% to 45.40% and reduces MAE from 0.9982 to 0.7822, showing that multi-peak modeling provides more effective

Table 4: Component ablation of CDF on CMU-MOSI. Bold values indicate the best performance. MPUM denotes Multi-Peak Uncertainty Modeling, and DSR denotes Degradation Source Rectification.

Model Variant	Components		CMU-MOSI					
	MPUM	DSR	Acc-7↑	Acc-5↑	Acc-2↑	F1↑	Corr↑	MAE↓
Base Model	×	×	33.83	37.26	78.39	78.13	0.6660	0.9982
+ MPUM	✓	×	45.40	53.54	84.02	84.11	0.7561	0.7822
+ DSR	✓	✓	**49.98**	**54.81**	**87.04**	**87.02**	**0.8042**	**0.6835**

Table 5: Ablation study of aleatoric and predictive uncertainty in MPUM on CMU-MOSI. Bold and underlined values denote the best and second-best results, respectively.

Model Variant	CMU-MOSI					
	Acc-7↑	Acc-5↑	Acc-2↑	F1↑	Corr↑	MAE↓
Only Predictive Uncertainty	<u>46.94</u>	<u>52.62</u>	<u>84.76</u>	<u>84.74</u>	0.7942	<u>0.7104</u>
Only Aleatoric Uncertainty	46.65	52.04	84.30	84.28	<u>0.7961</u>	0.7130
Aleatoric + Predictive (Full)	**49.98**	**54.81**	**87.04**	**87.02**	**0.8042**	**0.6835**

reliability estimates for dynamic fusion. Adding DSR further increases Acc_7 to 49.98% and $Corr$ to 0.8042, indicating that rectifying under-optimized representations improves uncertainty estimation. The cumulative gains confirm their complementarity.

4.2.2 Modality uncertainty rectification.

Table 3 reports modality-wise expected calibration error (ECE) on MVSA-Single, where lower values indicate better calibration. MPUM reduces text and image ECE from 25.98% and 27.45% to 20.46% and 17.69%, respectively. Although DSR alone does not improve calibration over the single-Gaussian baseline, combining it with MPUM further lowers ECE to 19.92% and 12.64%, yielding the best calibration for both modalities among the compared variants.

4.2.3 Uncertainty decomposition.

Table 5 shows that predictive uncertainty alone yields better classification results, while aleatoric uncertainty alone achieves slightly higher Corr. Combining both achieves the best performance across all six metrics, improving Acc-7 and F1 by 3.04 and 2.28 points over the predictive-only variant, respectively. These gains support their complementarity in dynamic modality weighting.

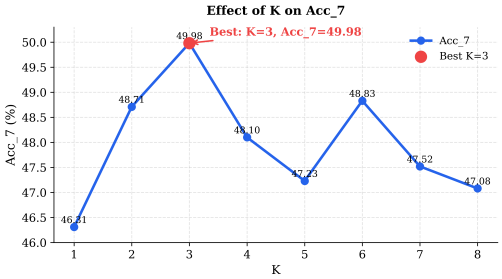
4.3 Further Analysis

4.3.1 Sensitivity to the number of components.

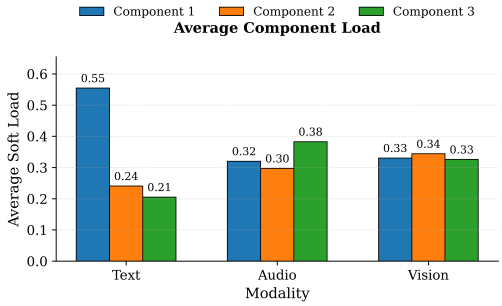
Figure 3(a) studies the effect of K. Increasing K from 1 to 3 raises Acc_7 from 46.31% to 49.98%, confirming that a single Gaussian is insufficient to represent multiple plausible sentiment patterns. Performance generally declines beyond $K = 3$, although $K = 6$ shows a local rebound, indicating that excessive components may fragment the predictive patterns and complicate routing. We therefore use $K = 3$ in all experiments.

4.3.2 Gaussian component utilization.

Figure 3(b) reports the dataset-level average mixture weights. All three components receive substantial loads, indicating that no component is globally neglected. Their utilization is also modality-dependent: text concentrates more strongly on Component 1, whereas audio and vision distribute their weights more evenly. Component 1 receives 55% for text. The other two components jointly retain 45%, preserving substantial contributions beyond the dominant component. Vision is nearly uniform. This difference suggests that the sample-conditional router learns modality-specific com-

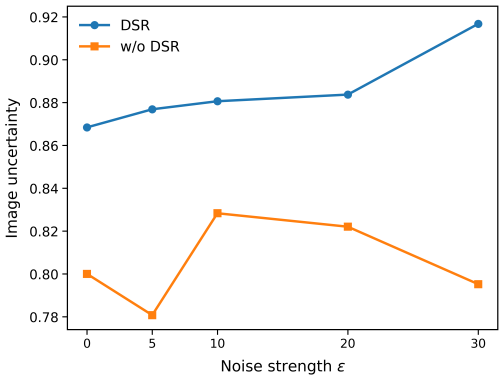


(a) Sensitivity of Acc_7 to the number of Gaussian components K on CMU-MOSI.

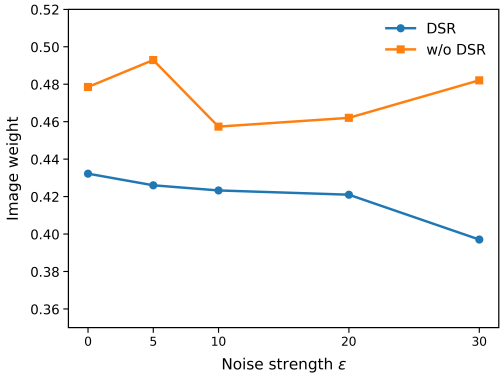


(b) Average mixture weights assigned to the three Gaussian components for each modality on CMU-MOSI.

Figure 3: Analysis of the Gaussian mixture components on CMU-MOSI.



(a) Estimated image uncertainty under increasing Gaussian noise.



(b) Dynamic image fusion weight under increasing Gaussian noise.

Figure 4: Image uncertainty and dynamic fusion weight under increasing image corruption on MVSA-Single.

binations of sentiment patterns rather than using a fixed mixture for all modalities. These results support retaining multiple components for modality uncertainty estimation.

4.3.3 Analysis of DSR.

DSR aims to reduce uncertainty bias arising from insufficient unimodal learning, so that dynamic weighting better reflects input reliability. To examine this intended effect, we progressively add Gaussian noise to a fixed test image on MVSA-Single while keeping its paired text unchanged. Increasing corruption should raise image uncertainty and reduce its fusion weight. As shown in Fig. 4(a)–(b), the model without DSR exhibits fluctuating uncertainty and weights, failing to consistently reflect the increasing corruption severity. With DSR, image uncertainty increases steadily while its weight decreases accordingly. These observations are consistent with DSR alleviating optimization-induced uncertainty bias: stronger unimodal learning provides a more reliable basis for assessing input degradation and adjusting modality contributions during fusion.

5 Conclusion

We propose CDF for robust multimodal sentiment analysis. DSR reduces optimization-induced uncertainty bias, while MPUM captures multi-pattern sentiment distributions with sample-conditional Gaussian mixtures. Experiments on three benchmarks verify its effectiveness and robustness.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript writing and language polishing, including improving grammar, readability, and clarity. All AI-assisted content was carefully reviewed and revised by the authors. The use of generative AI tools did not replace the authors' responsibility for the research design, experimental analysis, scientific claims, or final manuscript. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main paper and supplementary materials describe the CDF architecture, uncertainty estimation, training objectives and procedures, datasets, evaluation metrics, and ablation settings to support reproducibility. The source code will be publicly released upon acceptance to facilitate independent reproduction and verification of our results.

REFERENCES

Kamran Ali and Charles E. Hughes. A unified transformer-based network for multimodal emotion recognition, 2023. URL https://arxiv.org/abs/2308.14160.

Geetha A.V., Mala T., Priyanka D., and Uma E. Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Information Fusion*, 105:102218, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.102218. URL https://www.sciencedirect.com/science/article/pii/S1566253523005341.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. EMOE: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14314–14324, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fang_EMOE_Modality-Specific_Enhanced_Dynamic_Emotion_Experts_CVPR_2025_paper.html.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. *arXiv preprint arXiv:2410.04491*, 2024.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26876–26885, 2024.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413678. URL https://doi.org/10.1145/3394171.3413678.

Peng Hu, Yang Qin, Yuanbiao Gou, Yunfan Li, Mouxing Yang, and Xi Peng. Probabilistic multimodal learning with von mises-fisher distributions. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 5390–5398. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/600. URL https://doi.org/10.24963/ijcai.2025/600. Main Track.

Jingwang Huang, Jiang Zhong, Qin Lei, Jingpeng Gao, Yuming Yang, Sirui Wang, Peiguang Li, and Kaiwen Wei. Latent distribution decoupling: A probabilistic framework for uncertainty-aware multimodal emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. URL https://aclanthology.org/2025.findings-acl.1238/.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/huang22e.html.

Xun Jiang, Yufan Gu, Disen Hu, Yuqing Hou, Yazhou Yao, Fumin Shen, Heng Tao Shen, and Xing Xu. Multimodal learning on low-quality data with conformal predictive self-calibration, 2026. URL https://arxiv.org/abs/2605.03820.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation, 2023a. URL https://arxiv.org/abs/2308.07686.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition, 2023b. URL https://arxiv.org/abs/2303.13802.

Feng Liu, Ziwang Fu, Yunlong Wang, and Qijian Zheng. Tacfn: Transformer-based adaptive cross-modal fusion network for multimodal emotion recognition, 2025. URL https://arxiv.org/abs/2505.06536.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors, 2018. URL https://arxiv.org/abs/1806.00064.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive Modality Reinforcement for Human Multimodal Emotion Recognition from Unaligned Multimodal Sequences . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2554–2562, Los Alamitos, CA, USA, June 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.00258. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00258.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. Sentiment analysis on multi-view social data. In *MultiMedia Modeling*, pp. 15–27, 2016.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation, 2022. URL https://arxiv.org/abs/2203.15332.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research, 2020. URL https://arxiv.org/abs/2005.00357.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2359–2369, 2020. doi: 10.18653/v1/2020.acl-main.214. URL https://aclanthology.org/2020.acl-main.214/.

Shuo Sun, Haoyu Wang, Di Hu, Jiao Li, Yuejie Chen, and Zhihang Wang. Learning specific and shared representations for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26: 6579–6591, 2024. doi: 10.1109/TMM.2024.3353556.

Mani Kumar Tellamekala, Shahin Amiriparian, Björn W. Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition, 2023. URL https://arxiv.org/abs/2206.05833.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656/.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: Disentangled-language-focused multimodal sentiment analysis, 2025. URL https://arxiv.org/abs/2412.12225.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning, 2025a. URL https://arxiv.org/abs/2507.10203.

Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning, 2025b. URL https://arxiv.org/abs/2507.10213.

Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning, 2024. URL https://arxiv.org/abs/2407.09705.

Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning, 2024. URL https://arxiv.org/abs/2402.16897.

Shilin Xu, Yuan Sun, Xingfeng Li, Siyuan Duan, Zhenwen Ren, Zheng Liu, and Dezhong Peng. Noisy label calibration for multi-view classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):21797–21805, Apr. 2025. doi: 10.1609/aaai.v39i20.35485. URL https://ojs.aaai.org/index.php/AAAI/article/view/35485.

Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, and Lihua Zhang. Context de-confounded emotion recognition, 2023. URL https://arxiv.org/abs/2303.11921.

Dingkang Yang, Mingcheng Li, Xuecheng Wu, Zhaoyu Chen, Kaixun Jiang, Keliang Liu, Peng Zhai, and Lihua Zhang. Improving multimodal sentiment analysis via modality optimization and dynamic primary modality selection, 2026. URL https://arxiv.org/abs/2511.06328.

Yang Yang, Xixian Wu, and Qing-Yuan Jiang. Towards equilibrium: An instantaneous probe-and-rebalance multimodal learning approach. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 3552–3560. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/395. URL https://doi.org/10.24963/ijcai.2025/395. Main Track.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31 (6):82–88, 2016. doi: 10.1109/MIS.2016.94. URL https://ieeexplore.ieee.org/document/7742297.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis, 2017. URL https://arxiv.org/abs/1707.07250.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22123–22138, 2025. doi: 10.18653/v1/2025.acl-long.1075. URL https://aclanthology.org/2025.acl-long.1075/.

A Appendix

A.1 Datasets and Preprocessing

A.1.1 Datasets

We evaluate CDF on three widely used multimodal sentiment analysis benchmarks: CMU-MOSI, CMU-MOSEI, and MVSA-Single.

CMU-MOSI. CMU-MOSI contains 2,199 opinion utterances collected from online videos. Each utterance is annotated with a sentiment score ranging from -3 to 3 and contains textual, visual, and acoustic modalities. Following the standard split, we use 1,284 samples for training, 229 for validation, and 686 for testing.

CMU-MOSEI. CMU-MOSEI is a larger-scale multimodal sentiment dataset containing 22,856 video utterances. Each utterance is annotated with a sentiment score in $[-3, 3]$ and contains aligned textual, visual, and acoustic modalities. We follow the standard split of 16,326 training samples, 1,871 validation samples, and 4,659 test samples.

MVSA-Single. MVSA-Single contains image–text pairs collected from social media, where each sample is associated with a sentiment annotation. We use this dataset to evaluate the robustness of CDF under corrupted multimodal inputs.

A.1.2 Compared methods.

On CMU-MOSI and CMU-MOSEI, we compare CDF with Graph-MFN Zadeh et al. (2018), MulT Tsai et al. (2019), PMR Lv et al. (2021), MISA Hazarika et al. (2020), MAG-BERT Rahman et al. (2020), DMD Li et al. (2023b), DLF Wang et al. (2025), LDDU Huang et al. (2025), and MODS Yang et al. (2026). For MVSA-Single, we compare with EAU Gao et al. (2024), ECML Xu et al. (2024), NLC Xu et al. (2025), IPRM Yang et al. (2025), ARL Wei et al. (2025a), and CPSC Jiang et al. (2026). MOSI and MOSEI baselines are reproduced using official implementations with the same protocol, while CDF results are averaged over three runs. MVSA-Single results are adopted from CPSC under identical corruption settings.

A.1.3 Feature Representation

For CMU-MOSI and CMU-MOSEI, the textual modality is represented using 768-dimensional BERT-base-uncased features, while the visual and acoustic modalities use the publicly available pre-extracted features provided by the benchmarks. The three modalities are aligned at the utterance level and independently processed by modality-specific encoders.

No additional feature augmentation is introduced. The resulting modality representations are learned under the unimodal supervision and gradient-decoupled optimization of DSR, and are subsequently provided to MPUM for uncertainty estimation and dynamic fusion.

A.1.4 Corruption Protocol

Following CPSC, we evaluate robustness on MVSA-Single using Gaussian and salt-and-pepper corruption. For each corruption type, we consider two corruption strengths, denoted by $\epsilon \in \{5, 10\}$, in addition to the clean setting. The same corruption protocol is applied to all compared methods to ensure a fair comparison.

A.2 Implementation Details

We implement CDF using PyTorch. All experiments are conducted with random seeds 1111, 1112, and 1113, and the best checkpoint of each run is selected according to validation performance. The complete optimization, architecture, MPUM, dynamic-weighting, and loss settings are summarized in Tables 6 and 7.

Table 6: Optimization and architecture settings on CMU-MOSI and CMU-MOSEI.

Hyperparameter	**CMU-MOSI**	**CMU-MOSEI**
Optimization settings		
Random seed	1111/1112/1113	1111/1112/1113
Optimizer	AdamW	AdamW
Main learning rate	3×10^{-5}	8×10^{-5}
BERT learning rate	1×10^{-5}	1×10^{-5}
Weight decay	0.05	0.05
Batch size	32	32
Maximum epochs	100	100
Learning-rate scheduler	Warmup cosine	Warmup cosine
Warmup epochs	2	2
Minimum learning-rate ratio	0.1	0.1
Gradient clipping	1.0	1.0
Architecture settings		
Text backbone	BERT-base-uncased	BERT-base-uncased
Input feature dimensions (L/A/V)	$768/20/5$	$768/35/74$
Maximum sequence lengths (L/A/V)	$50/500/375$	$50/500/500$
Hidden dimension	128	128
Feature encoder depth	2	3
Projection depth	2	3
Cross-modal attention layers	1	2
Number of attention heads	4	4
Attention dropout	0.4	0.4
Number of proxy tokens	8	8

Table 7: MPUM, dynamic-weighting, and loss settings on CMU-MOSI and CMU-MOSEI.

Hyperparameter	**CMU-MOSI**	**CMU-MOSEI**
MPUM settings		
Number of Gaussian components K	3	3
Entmax order α_{ent}	1.5	1.5
Component variance range	$[10^{-4}, 9]$	$[10^{-4}, 9]$
Final router temperature (L/A/V)	$0.6/0.7/0.7$	$0.6/0.7/0.7$
Uncertainty and dynamic-weighting settings		
Aleatoric coefficient λ_a	0.7	0.7
Predictive coefficient λ_p	0.3	0.3
Routing-ambiguity coefficient γ	1.0	1.0
Weight-mapping exponent β	1.0	1.0
Dynamic-weight strength	0.5	1.0
Equal-weight warmup epochs	10	10
Dynamic-weight ramp epochs	10	10
Loss settings		
Task loss	MSE	MSE
DSR unimodal coefficient α	4.0	4.0
Label-smoothing coefficient	0.02	0.10
GMM NLL coefficient λ_{nll}	0.1	0.1
Routing-consistency coefficient λ_{route}	1×10^{-3}	1×10^{-3}
MPUM regularization warmup/ramp epochs	$5/10$	$5/10$

Optimization settings. All models are optimized using AdamW for a maximum of 100 epochs. Separate learning rates are used for BERT and the remaining model parameters. Specifically, the main learning rates are set to 3×10^{-5} and 8×10^{-5} on CMU-MOSI and CMU-MOSEI, respectively, while the BERT learning rate is fixed to 1×10^{-5}. We adopt a warmup cosine scheduler with a two-epoch warmup stage and a minimum learning-rate ratio of 0.1. The weight decay, batch size, and maximum gradient norm are set to 0.05, 32, and 1.0, respectively.

Table 8: Comparison of gradient-decoupling strategies in DSR on CMU-MOSI.

Model Variant	CMU-MOSI					
	Acc-7↑	Acc-5↑	Acc-2↑	F1↑	Corr↑	MAE↓
w/o DSR	45.40	53.54	84.02	84.11	0.7561	0.7822
Unimodal Supervision (US)	<u>47.36</u>	<u>53.61</u>	<u>85.67</u>	85.61	0.7888	0.7096
Partial Decoupling (PD)	46.94	53.06	83.67	<u>85.82</u>	<u>0.7976</u>	<u>0.7072</u>
Staged Decoupling (SD)	46.21	52.04	83.67	85.52	0.7970	0.7137
Ours	**49.98**	**54.81**	**87.04**	**87.02**	**0.8042**	**0.6835**

Notes: Bold and underlined values denote the best and second-best results, respectively. "w/o DSR" removes both unimodal supervision and gradient decoupling. "US" introduces direct unimodal supervision while allowing full fusion gradients to update the modality encoders. "PD" scales the fusion gradients by 0.5 throughout training. "SD" blocks fusion gradients during the first 50 epochs and restores them during the remaining 50 epochs. "Ours" blocks fusion gradients from updating the modality encoders throughout training.

Architecture settings. BERT-base-uncased is used as the textual backbone. The input feature dimensions of the language, acoustic, and visual modalities are $768/20/5$ on CMU-MOSI and $768/35/74$ on CMU-MOSEI. All modalities are projected into a shared 128-dimensional latent space. For CMU-MOSI, the feature encoder, projection module, and cross-modal attention module contain two, two, and one layers, respectively; the corresponding depths on CMU-MOSEI are three, three, and two. The cross-modal attention module uses four attention heads with a dropout rate of 0.4. The proxy representation is constructed using eight proxy tokens.

MPUM and dynamic weighting. MPUM models the sample-conditional sentiment distribution of each modality using three Gaussian components. The sparse router adopts $\mathrm{Entmax}_{1.5}$, and the component variances are constrained to $[10^{-4}, 9]$. The final router temperatures for the language, acoustic, and visual modalities are $0.6/0.7/0.7$.

Aleatoric and predictive uncertainties are combined using $(\lambda_a, \lambda_p) = (0.7, 0.3)$ on both CMU-MOSI and CMU-MOSEI. The routing-ambiguity coefficient γ and weight-mapping exponent β are both set to 1.0. The dynamic-weight strengths are set to 0.5 and 1.0 on CMU-MOSI and CMU-MOSEI, respectively. To stabilize uncertainty-guided fusion, equal modality weights are used during the first 10 epochs, followed by a 10-epoch transition to dynamic weighting.

Loss settings. The multimodal and unimodal sentiment prediction objectives are implemented using mean squared error. The DSR unimodal-loss coefficient α is set to 4.0. The label-smoothing coefficients are set to 0.02 and 0.10 on CMU-MOSI and CMU-MOSEI, respectively. The GMM negative log-likelihood and routing-consistency losses are weighted by $\lambda_{\mathrm{nll}} = 0.1$ and $\lambda_{\mathrm{route}} = 10^{-3}$, respectively. The MPUM regularization objectives are activated after a five-epoch warmup and linearly increased during the subsequent 10 epochs.

A.3 Additional Ablation Studies

A.3.1 Sensitivity to Uncertainty Weights

We examine how the relative contributions of aleatoric and predictive uncertainty affect sentiment prediction. We vary the aleatoric coefficient λ_a over $\{0.5, 0.6, 0.7, 0.8\}$ and set $\lambda_p = 1 - \lambda_a$, evaluating Acc-2 on CMU-MOSI.

As shown in Fig. **??**, the adopted setting $(0.7, 0.3)$ achieves the highest Acc-2 among the tested configurations at 87.04%, compared with 86.73%, 86.24%, and 85.42% for $(0.5, 0.5)$, $(0.6, 0.4)$, and $(0.8, 0.2)$, respectively. Performance does not improve monotonically with the aleatoric coefficient: increasing it from 0.7 to 0.8 reduces Acc-2 by 1.62 percentage points. This decline suggests that excessively reducing the contribution of cross-pattern predictive disagreement can weaken reliability-based fusion. The results provide empirical support for the adopted balance between within-pattern dispersion and cross-pattern disagreement.

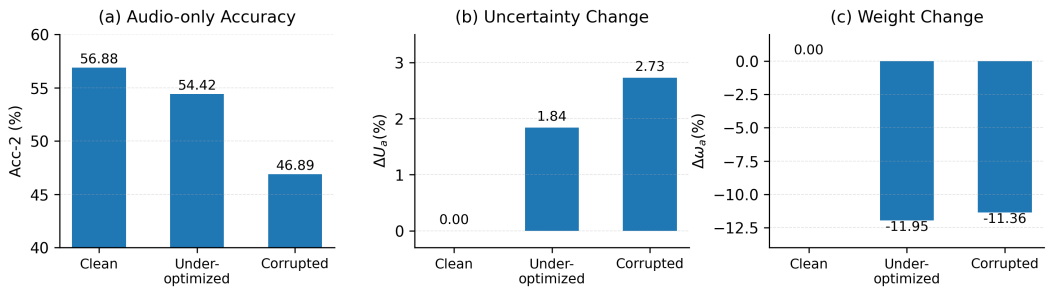


Figure 5: Controlled analysis of the acoustic modality on CMU-MOSI under clean, under-optimized, and input-corrupted conditions. (a) Audio-only Acc-2; (b) relative uncertainty change with respect to the clean condition; and (c) relative dynamic-weight change.

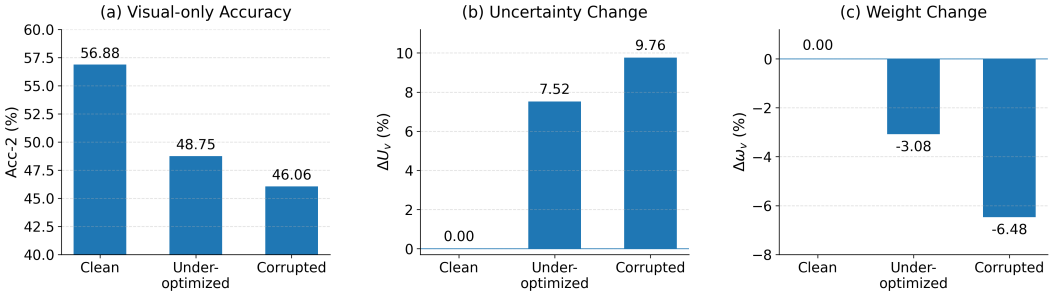


Figure 6: Controlled analysis of the visual modality on CMU-MOSI under clean, under-optimized, and input-corrupted conditions. (a) Visual-only Acc-2; (b) relative uncertainty change with respect to the clean condition; and (c) relative dynamic-weight change.

A.3.2 Gradient-Decoupling Strategies

As shown in Table 8, US improves all six metrics over the baseline without DSR, demonstrating the benefit of direct unimodal supervision. PD and SD yield mixed results: both achieve higher Corr than US but lower Acc-7, Acc-5, and Acc-2, suggesting that partial or staged decoupling does not consistently improve performance. Full decoupling achieves the best results across all metrics, including 49.98% Acc-7, 0.8042 Corr, and 0.6835 MAE. Compared with US, it improves Acc-7 by 2.62 percentage points and reduces MAE by 0.0261, supporting the contribution of optimization-path design beyond additional supervision. These results favor combining direct unimodal supervision with persistent gradient separation. This design prevents fusion gradients from directly updating modality encoders, while allowing their representations to participate in cross-modal interaction during the forward pass.

A.3.3 Controlled Analysis of Optimization- and Input-Induced Degradation

To directly examine the confounding problem discussed in the Introduction, we construct three controlled conditions for the acoustic and visual modalities. The *clean* condition uses normally optimized encoders and uncorrupted inputs. The *under-optimized* condition retains clean inputs but uses an insufficiently optimized modality encoder, isolating representation degradation caused by inadequate learning. In contrast, the *corrupted* condition retains normal optimization but injects external noise into the modality input, representing input-dependent unreliability. We report unimodal Acc-2 and the relative changes in estimated uncertainty and dynamic weight with respect to the clean condition.

Analysis. Figures 5 and 6 show that optimization-induced representation weakness and input corruption trigger similar uncertainty-weighting responses. For audio, under-optimization reduces Acc-2 from 56.88% to 54.42%, increases uncertainty by 1.84%, and decreases its weight by 11.95%; corruption further lowers Acc-2 to 46.89% and increases uncertainty by 2.73%. A clearer monotonic

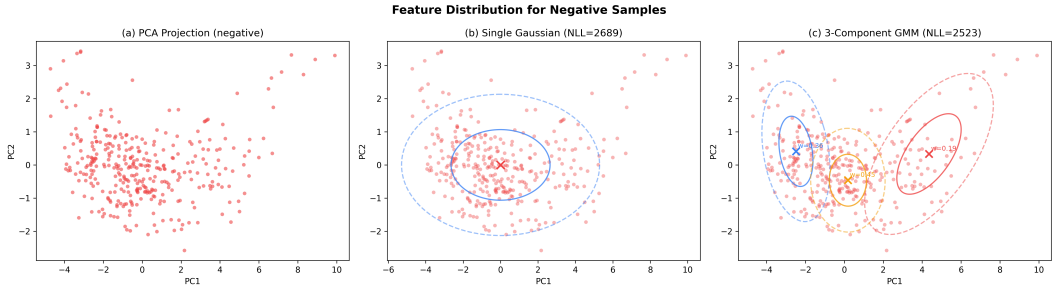


Figure 7: PCA visualization and probabilistic fitting of textual modality representations for negative samples from CMU-MOSI. Compared with the single Gaussian, the three-component GMM captures multiple local patterns and reduces the negative log-likelihood from 2689 to 2523.

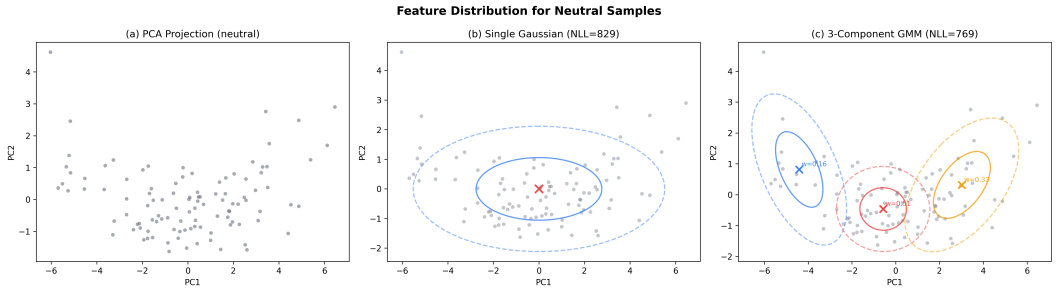


Figure 8: PCA visualization and probabilistic fitting of textual modality representations for neutral samples from CMU-MOSI. Compared with the single Gaussian, the three-component GMM captures several locally concentrated patterns and reduces the negative log-likelihood from 829 to 769.

trend appears for vision: Acc-2 decreases from 56.88% to 48.75% and 46.06%, while uncertainty increases by 7.52% and 9.76% and the corresponding weights decrease by 3.08% and 6.48%.

These consistent trends validate the motivation of DSR: an under-optimized modality can exhibit the same high-uncertainty and low-weight behavior as a genuinely corrupted modality, causing optimization artifacts to be confused with input unreliability. The acoustic weights are not strictly ordered because each modality weight is normalized relative to the other modalities. Overall, the results support reducing optimization-induced uncertainty bias before dynamic weighting, rather than interpreting all representation degradation as input corruption.

A.3.4 Additional Visualization of Multi-Peak Textual Patterns

To complement the positive-sample visualization in the main paper, we further compare single-Gaussian and three-component GMM fitting on the textual modality representations of negative and neutral samples from CMU-MOSI.

Analysis. Figures 7 and 8 extend the positive-sample analysis in the main paper. For both sentiment categories, the single Gaussian covers the representations using one coarse global distribution, whereas the three-component GMM captures several locally concentrated patterns and achieves lower negative log-likelihood. Together with the positive-sample results, these observations indicate that textual representations exhibit heterogeneous local structures across sentiment categories, providing empirical motivation for multi-component uncertainty modeling.